View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Projetos e Dissertações em Sistemas de Informação e Gestão do Conhecimento

DOUGLAS VIEIRA SANTOS

PREDIÇÃO DE LINKS EM REDES DE COAUTORIA: UM ESTUDO UTILIZANDO A TEORIA DA EVOLUÇÃO ESPECTRAL EM REDES COMPLEXAS

BELO HORIZONTE

2014

DOUGLAS VIEIRA SANTOS

PREDIÇÃO DE LINKS EM REDES DE COAUTORIA: UM ESTUDO UTILIZANDO A TEORIA DA EVOLUÇÃO ESPECTRAL EM REDES COMPLEXAS

Projeto de Dissertação de Mestrado apresentado à Universidade FUMEC como requisito parcial para obtenção do título de Mestre em Sistemas de Informação e Gestão do Conhecimento.

UNIVERSIDADE FUMEC

FACULDADE DE CIÊNCIAS EMPRESARIAIS

MESTRADO PROFISSIONAL EM SISTEMAS DE INFORMAÇÃO E GESTÃO DO CONHECIMENTO

Orientador: Dr. ORLANDO ABREU GOMES

Coorientador: Dr. FERNANDO SILVA PARREIRAS

BELO HORIZONTE

2014

Resumo

SANTOS, Douglas Vieira. Predição de Links em Redes de Coautoria: um estudo utilizando a teoria da evolução espectral em redes complexas. 2014. Projeto de Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento) – Faculdade de Ciências Empresariais, Universidade FUMEC, Belo Horizonte, 2014.

Este Projeto de Dissertação de Mestrado pretende apresentar as diretrizes para realizar a pesquisa sobre predição de links em redes de coautoria. Por meio de um software a ser desenvolvido, considerando técnicas de álgebra linear na evolução das redes representadas matematicamente por grafos, uma rede de coautoria que evolui ao longo do tempo será estudada. O resultado esperado do trabalho é a validação do software ao mostrar que a predição de links foi realizada. Os resultados serão comparados a outras técnicas definidas por meio de uma revisão bibliográfica.

Palavras-chave: Predição de Links, Evolução Espectral, Redes de Coautoria, Grafos

Abstract

This dissertation project aims to present the guidelines to carry out research on predicting the links in co-authorship networks. Through a software to be developed, considering linear algebra techniques in the evolution of networks represented mathematically by graphs, a network of co-authorship that evolves over time will be studied. The expected result of the work is the validation of the software to show that the prediction of links was performed. The results are compared to other techniques defined by means of a literature review.

Keywords: Link Prediction, Spectral Evolution, Co-authorship Networks, Graphs.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Por volta de 300 a.C, o matemático grego Euclides estabeleceu que a menor distância entre dois pontos em um espaço bidimensional é um segmento de reta (CALLIOLI; DOMINGUES; COSTA, 2007). Ao considerar os dois pontos como nós e o segmento de reta como uma interligação tem-se uma rede com dois nós e uma aresta.

Figura 1 – Rede contendo dois nós e uma aresta

Toda rede possui características fundamentais para compreensão de sua estrutura. Não apenas deve-se considerar que a rede possui nós interligadas. Para exemplificar, é comum pensar em um conjunto de pessoas que são amigas. Ao longo da vida as pessoas vão fazendo novos amigos o que aumenta o conjunto. A inteligação entre as pessoas é o laço de amizade e as pessoas são os nós. A rede de amizade entre pessoas tende a crescer a medida que novos nós (pessoas) entram na rede e geram as arestas, as relações de amizade, o que permiti supor uma evolução da rede ao longo do tempo. Em uma rede de colaboração entre autores de artigos científicos, a característica da evolução também pode ser apontada. Ao longo do tempo novos trabalhos são produzidos em conjunto por dois ou mais autores, criando uma interligação entre eles.

Kunegis, Fay e Bauckhage (2010) propõe um modelo matemático que ultiliza algebra linear para explicar a evolução redes: o modelo da evolução espectral.

Uma maneira de representar uma rede de forma que possa passar por um estudo utilizando um formalismo matemático é construindo seu grafo. Um grafo é um desenho que reperesenta os nós e as interligações (arestas) entre eles. A estratégia utilizada por Kunegis é desenhar a matriz de adjacência desse grafo e apresentar o crescimento da rede pela decomposição dessa matriz em seus autovalores e autovetores. Para Kunegis a evolução da rede acontece dentro do mesmo subespeço vetorial ao considerar que os autovetores permanecem aproximadamente constantes enquanto os autovalores variam.

Ao compreender que a teoria de Kunegis (KUNEGIS, 2011) possibilita analisar a evolução temporal de uma rede, pretende-se prever o surgimento de novas interligações entre os nós: os links.

O projeto está escrito com a seguinte estrutura: na seção 1.1 é apresentado o problema a ser resolvido, na seção 1.2 apresenta-se a justificativa, e na seção 1.3 mostra-se os objetivos do trabalho. O capítulo 2 mostra a revisão da literatura realizada com o intuito de encontrar trabalhos relacionados e fornecer embasamento teórico para o projeto. No capítulo 3 mostra-se a metodologia da pesquisa e o que será executado para se atingir os objetivos e no capítulo 4 apresenta-se do cronograma do projeto.

1.1 Problema

A rede de coautoria da *Plataforma Lattes* (http://lattes.cnpq.br) é uma rede composta por uma grande quantidade de pesquisdores. Pode-se afirmar que a rede não é estática pois novas colaborações entre os pesquisadores estão surgindo ao longo do tempo. Considerando a teoria da evolução espectral das redes proposta por Kunegis (2011), qual o desempenho ao aplica-la para realizar a predição de links na rede de coautoria dos pesquisadores da *Plataforma Lattes*?

1.2 Justificativa

Em 2003 Liben-Nowell et al. apresentam uma análise do desempenho das técnicas de predição de links baseadas em vizinhança comum e a distância dos nós nos grafos (LIBEN-NOWELL; KLEINBERG, 2003). Ao considerar a técnica mais simples, observa-se a baixa performance quando se considera que um link que surge randomicamente em uma rede está correto. Uma maneira de realizar esse experimento é considerar uma rede fictícia com o mesmo número de nós de uma rede real. Por uma simulação computacional é possível simular aonde irá surgir o próximo link considerando a proximidade dos nós. Possuindo o resultado da simulação, verifica-se a acurácia do surgimento do link na evolução da rede real. A performance ficou abaixo de 0,5 %, ou seja, 99,5 % dos links randômicos não corresponderam aos links reais.

predictor	astro-ph	cond-mat	gr-qc	hep-ph	hep-th
probability that a random prediction is correct	0.475%	0.147%	0.341%	0.207%	0.153%

Figura 2 – Probabilidade de um link que surge randomicamente em uma das redes de coautoria analisadas estar correto

FONTE: (LIBEN-NOWELL; KLEINBERG, 2003)

A figura 2 apresenta as redes de coautoria, *astro-ph* uma categoria de publicação de trabalhos sobre astrofísica, *cond-mat* uma categoria de publicação de trabalhos sobre matéria condesada, *gr-qc* uma categoria de publicação de trabalhos sobre relatividade geral e cosmologia quântica, *help-ph* uma categoria de publicação de trabalhos sobre a fenomenologia da física de alta energia , *hep-th* uma categoria de publicação de trabalhos sobre a teoria da física de alta energia . Essas categorias de publicações são mostradas no sítio da biblioteca virtual da Universidade de Cornell, o repositório é denominado ArXiv (http://arxiv.org/). Em todas as redes resultantes da colaboração entre os autores em suas respectivas categorias, observa-se o baixo desenpenho da técnica. A tarefa de prever links com alta probabilidade de acerto apresenta-se como um desafio.

O artigo (KUNEGIS; FAY; BAUCKHAGE, 2010) apresenta uma análise da evolução da rede ao longo do tempo onde o foco do estudo é a estrutura das redes e não características de vizinhaça ou distância entre nós. Avaliar o método na rede da *Plataforma Lattes* pode ser útil para a construção do mapeamento de áreas de pesquisas no Brasil, ao se prever o crescimento e até mesmo o surgimento de áreas de pesquisa.

A construção de um software para se prever links pode possibilitar a otimização dos sistemas de recomendações para sugestões de novos amigos em uma rede social e até mesmo sugestões para compra produtos por pessoas que possuem perfis comuns de acordo com as característcas de compras em lojas virtuais.

1.3 Objetivo Geral

Avaliar a predição de links na rede de coautoria da *Plataforma Lattes* apropriando-se da tecnologia da evolução espectral proposta por Kunegis (2011).

1.4 Objetivos Específicos

1- Desenvolver um software para gerar a matriz de adjacência do grafo da rede da plataforma Lattes.

2- Montar o mapa de interligações dos autores de trabalhos.

3- Apontar aonde está ocorrendo a evolução na rede analisada.

4- Apontar o provável surgimento de novas colaborações entre os autores.

5- Desenvolver um software que poderá ser aplicado em outros modelos de redes que evoluam ao longo do tempo para avaliar a preedição de links.

2 A Revisão da Literatura

2.1 Trabalhos Relacionados

Para encontrar trabalhos relacionados à temática do projeto, foi executado o procedimento de revisão da literatura construído por Kitchenham (2004). O processo consiste em escolher bases repositórios de artigos científcos e realizar uma busca baseada em uma única chave de pesquisa pertinente a todas as bases escolhidas. A revisão da literatura se iniciou pela busca de trabalhos nas seguintes bases de repositórios de trabalhos científicos: *ACM Digital Library* (http://dl.acm.org), *IEEExplorer* (http://ieee.org) e *Science Direct* (http://sciencedirect.com).

Foram encontrados 199 trabalhos. Após passarem por um processo de filtragem que tomou por critérios de inclusão: artigos publicados em jornais e revistas, artigos publicados em anais de congressos; e critérios de exclusão: livros e manuais; juntamento com a leitura dos resumos dos trabahos, o número passou para 139. Esses trabalhos passaram por um processo de extração de dados que consiste na leitura dos mesmos buscando sempre as mesmas características em comum: apresentação de técnicas para a predição de links, procedimentos executados para se realizar a predição e exemplos de análises de redes que evoluem ao longo do tempo. Ao final da revisão restaram 103 trabalhos.

Os trabalhos resultantes passaram por um processo de estudo de similaridade do conteúdo. O intuito de realizar esse estudo foi de encontrar quais são os trabalhos mais referenciados quando se estuda predição de links. O outro objetivo foi encontrar a quantidade de estudos que estão utilizando a teoria da evolução espectral para realizar a predição de links. A estratégia escolhida para mostrar o resultado foi desenhar um grafo da rede de similaridade dos trabalhos. Nesse grafo os nós representam um trabalho e as arestas estão de forma direcional mostrado o laço de similaridade existente. Depois de minerar os dados em uma planilha eletrônica o grafo da rede foi construído.

Todos os trabalhos foram ordenados com um único ID (número de identificação) e o resultado é mostrado na figura 3.

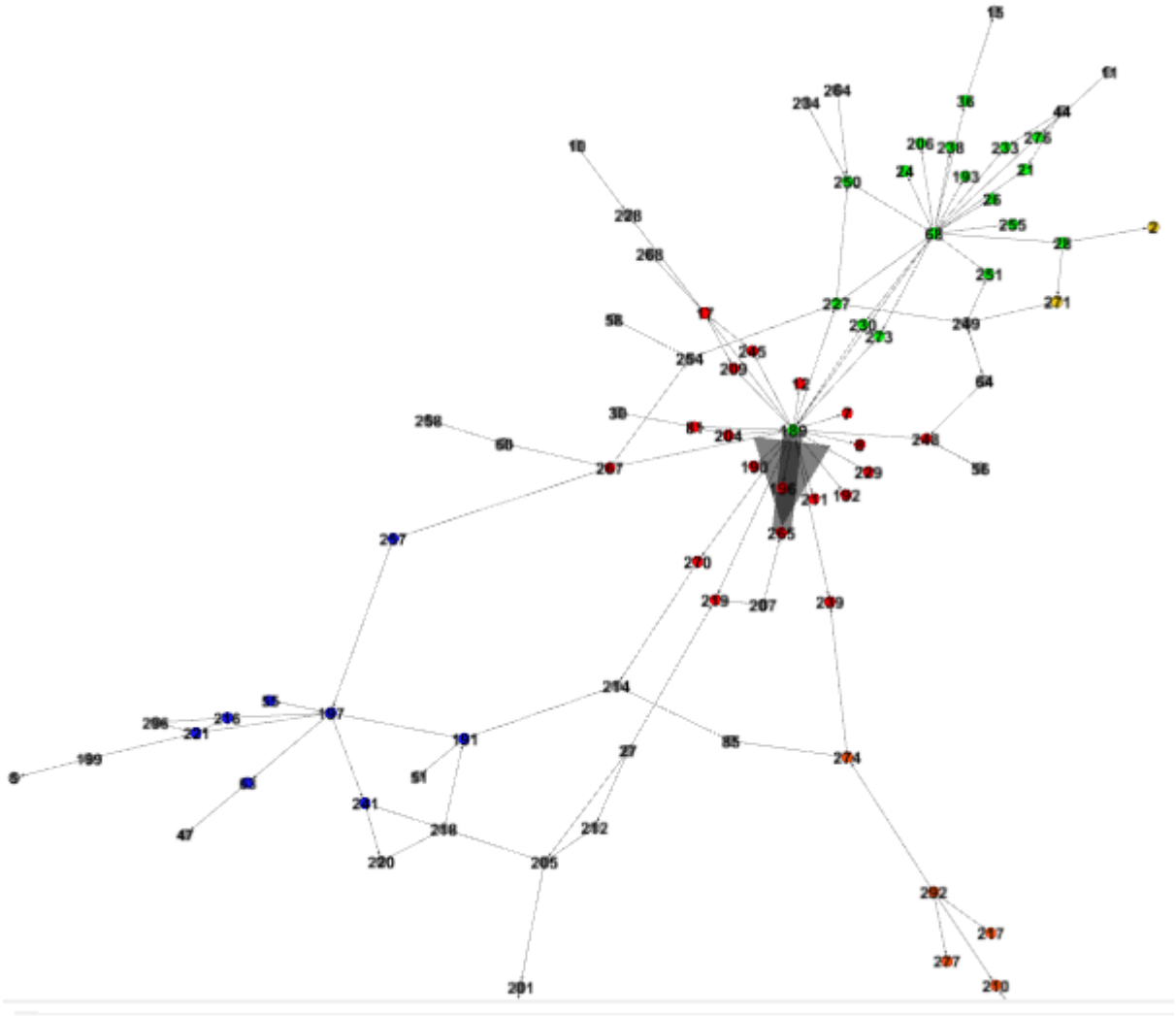

Figura 3 – Rede de similiridade dos trabalhos resultantes do processo de revisão sistemática da literatura

A rede foi construída da forma direcional considerando a data de publicação do trabalho. Exemplo: um trabalho publicado em 2006 é similar a um trabalho publicado em 2005, logo a direção é dada como partindo de 2005 para chegar em 2006.

Outro obejtivo da revisão é obter uma classificação dos trabalhos. A classificação é resultante do processo de leitura dos trabalhos onde o autor de cada um descreve o método teórico que está sendo utilizado em sua pesquisa. Os trabalhos foram agrupados e as categorias encontradas foram:

1 - Estacionário Probabilístico: consiste em analisar a rede sem considerar a evolução ao longo do tempo. Em um ponto estacionário da rede é aplicado um algorítimo para identificar a probabilidade do surgimento de um próximo link.

2 - Fatores Temporais: a análise da rede considera fatores que ocorrem ao longo do tempo que influênciam na a geração de links. Não são considerados a estrutura da rede

nem as características dos nós, os fatores temporais são fatores históricos como relatado por Tylenda, Angelova e Bedathur (2009).

3 - Força de Ligação: O método considera a característica do nó em atrair mais nós para se conectarem a ele. Chang e Yao (2011), apresentam o modelo denominado AF (força de atração), que considera a característica da força de ligação para apontar onde os novos links irão surgir na rede.

4 - Agrupamentos baseados em eventos: Dentro das redes pode-se apontar subredes que são os agrupametos de nós em torno de uma característica em comum. Ao analisarem a rede profissional do "Linkedin", Rodriguez e Rogati (2012) mostram que a ocorrência de um evento, como uma conferência de determinado grupo de profissioais, gera novas conexões entre os participantes dentro da rede profissional do *Linkedin* (http://linkedin.com).

5 - Similaridade dos nós: a similaridade dos nós considera características externas à rede nas quais os nós são comuns. Se duas pessoas participam de uma rede social e as duas tem um gosto musical semelhante, uma sugestão para que as duas pessoas formem um link é pertinente. Makrehchi (2011), chamam as características externas de "Hiden Topic", ao apresentarem seu artigo *Social link recommendation by learning hidden topics.*

6 - Proximidade, vizinhança comum: a categoria que agrupa a maior parte dos trabalhos encontrados. Consiste em dizer qual será o próximo link basedo nas ligações em comum existentes pelos nós da rede. Se um nó "A"é conectado a um nó "B"e um nó "C"tambem é conectado a "B", sugere-se que a novo link surgirá entre "A"e "C". Um exemplo de trabalho dentro dessa categoria é o publicado Dong et al. (2011), que considera as interações de vizinhaça comum estrapolando as ligações mais próximas dos nós considerando as ligações de maior grau de separação.

A ferramenta de análise de grafos *Gephi 8.2* (https://gephi.org), possibilita utilizar o algorítimo *Yufan Hu*, que centraliza os nós de acordo com a sua força de interação. Baseado na força de interação dos nós e na similaridade utilizada para construção do grafo, os nós foram coloridos na figura 3 para representar os nós que pertecem a um grupo. As cores são: vermelho, verde, laranja, azul e amarelo. Os nós "pilares"representam os principais trabalhos associados à temática do projeto, os principais são:

1 - ID 189, Kashima e Abe (2006), apresentam um modelo probabilístico para descrever evolução de uma rede biológica. É o trabalho pilar para todos nos quais os nós estão coloridos de vermelho na figura 3.

2 - ID 68, Liben-Nowell e Kleinberg (2003), apresentam técnicas utilizadas para prever links e suas performances ao serem aplicadas em redes de coautoria. É o trabalho pilar para todos nos quais os nós estão coloridos de verde na figura 3.

3 - ID 292, Lü e Zhou (2011), apresentam o progresso das tecnologias desenvolvidas até o ano de 2011. São citados 175 trabalhos nas referências bibliográficas do artigo. É o

trabalho pilar para todos nos quais os nós estão coloridos de laranja na figura 3.

4 - ID 197, Ahmad et al. (2010), apresenta uma análise considerando os nós participando de múltiplas redes. A predição de links é dominida "Internetwork Prediction". É o trabalho pilar para todos nos quais os nós estão coloridos de azul na figura 3.

A teoria da evolução espectral proposta por Kunegis, Fay e Bauckhage (2010), é representada na figura 3 pelos trabalhos: (KUNEGIS; LOMMATZSCH, 2009), (KUNEGIS; FAY; BAUCKHAGE, 2010), (ALLALI; MAGNIEN; LATAPY, 2011) que estão coloridos de amarelo. Essa teoria pode ser considerada uma categoria, pois é o único método encontrado na revisão que utiliza métodos baseados em álgebra linear para explorar a estrutura da rede. Ainda de acordo com a figura 3, conclui-se que a teoria da evolução espectral está sendo pouco explorada, e o presente projeto será pioneiro em utiliza-la para prever links em uma rede de coautoria brasileira, a *"Plataforma Lattes"*.

A inspiração para o desenvolvimento do projeto é a tese de doutorado de Kunegis (2011) entitulada: *On the Spectral Evolution of Large Networks*. Este projeto de dissertação de mestrado tem sua pesquisa direcionada para o modelo aplicado, o que não foi encontrado na revisão da literatura. Não é o objetivo desenvolver ou aprimorar uma técnica de predição de links. O trabalho está em avaliar o desempenho da técnica de Kunegis na rede de coautoria da *"Plataforma Lattes"*.

2.2 Redes

Redes estão em todos os lugares. Segundo SANTANA (2004), o conceito de redes é oriundo do latin *retíolos* (conjunto de linhas entrelassadas). A especificidade de interesses na ciência tem direcionado o estudo das redes. FONSECA e ONeill (2001) aborda o conceito de redes como um entrelaçamento de nós e fios.

Quando se fala em mineração de dados, o olhar para as redes torna-se analítico no sentido de desenvolver ferramentas (KUNEGIS; FAY; BAUCKHAGE, 2010). Um exemplo são as máquinas de buscas que utilizam sistemas de recomendações. Tais sistemas se valem de propriedades que caracterizam os modelos das redes. Kunegis, Fay e Bauckhage (2010) reforçam que as redes podem ser classificadas em três tipos de propriedades: estruturas das interligações, tipos de interligações e metadados.

Estrutura das interligações: Links podem ser sem direções definidas, direcionados ou bipartidos. Os nós de uma rede podem ser interligados de forma que a direção da interligação não tenha importancia. Se A está ligado a B, B está ligado a A. Levando em conta a importância da direção da interligação, considera-se que C pode estar conectado a D mas D pode não estar conectado a C. Uma rede bipartida pode ser a base de artigos dos encontros anuais da AMPAD (Associação Nacional de Pós-Graduação e Pesquisa em

Administração), onde existem dois tipos de nós; autores e publicações. Os autores são interligados por suas publicações produzidas em conjunto.

Tipos de Interligações: simples, múltiplas, sem peso ou com peso. A ligação simples pode ser exemplificada por uma relação de amizade existente em uma rede social na internet, o *"Facebook"* (http://facebook.com). Ao se estabelecer uma relação de amizade dentro do contexto da rede social *Facebook*, o link permanece constante sem que as duas pessoas consigam mostrar o quão forte é amizade entre elas. Assim o link é classificado como "sem peso". No caso de ligações múltiplas, um exemplo é uma mensagem enviada de uma pessoa para outra, as mensagens podem representar a interligação e quanto mais mensagens forem enviadas maior o número de ligações. As ligações com peso podem ser exemplificadas em uma rede de comércio internacional. Os países são os nós da rede e as relações comerciais de importação e exportação representam as interligações. O Petróleo seria um exemplo onde o volume exportado de um país para outro seria o fator de peso na interligação.

Metadados: De uma forma geral redes são criadas pelo processo de crescimento em que nós e conexões são adicionadas a estrutura ao longo do tempo. Para Kunegis, Fay e Bauckhage (2010) a rede no inicío de sua construção evolui em várias direções e se fosse traduzir a evolução em espaços vetorias estes estaria representando diferentes bases de subespaços. Quando a rede já está muito complexa, contendo vários nós e vários links, o subespaço vetorial representando a direção do crescimento da rede, permanece aproximadamente constante.

2.3 Redes de Coautoria

Na comunicação científica, o intercâmbio têm facilitado as relações entre autores e as áreas do conhecimento, contribuindo para experiências interdisciplinares sólidas (SILVA; BARBOSA; DUARTE, 2012). Artigos científicos gerealmente são escritos por mais de um autor. Considerando o artigo como o elemento de interligação entre os autores, tem-se uma rede de coautoria entre eles.

Para Newman (2001), as redes de coautoria talvez sejam mais genuínas do que algumas redes sociais. Em redes de coautoria, os autores de um artigo científico se conhecem antes de realizarem um trabalho juntos. Em redes sociais uma pessoa pode nunca ter visto a outra e estabelecer uma relação de amizade.

Segundo Brandão, Parreiras e Silva (2007), nas redes de coautoria os autores de trabalhos publicados em veículos de difusão do conhecimento são considerados atores, sendo que para cada autoria em conjunto entre dois atores é criado um laço relacional entre eles. No estudo das redes de coautria Parreiras et al. (2006), ao avaliarem as métricas da rede de coautoria no campo de ciência da informação no Brasil, apontaram a

existência de uma baixa colaboração global entre os autores. Nesse estudo Parreiras e seus colaboradores identificaram que as novas colaborações estão surgindo dentro de grupos restritos dentro da rede, denominados subredes. A colaboração em um âmbito geral da rede não foi detectada.

Os fatores de crescimento de uma rede de coaturia, bem como a previsão de novos links tem sido objeto de estudos. Sun et al. (2011) relizaram um trabalho onde propõe uma técnica para prever as relações de coautoria em redes bibliográficas heterogêneas, compostas por artigos, citações, lviros entre outros trabalhos.

2.4 Predição de Links

Prever Links em uma rede é apontar onde novas conexões irão surgir. Segundo Papadimitriou, Symeonidis e Manolopoulos (2011), redes sociais como o "Facebook.com", "My space", "His.com", contêm gigabytes de dados que podem ser manipulados para se realizar as previsões de novas conexões. As máquinas de buscas dessas redes levam em consideração a quantidade de amigos em comum que duas pessoas possuem para sugerir uma nova conexão. Os sistemas de sugestão levam em consideração o grau de separação. Para entender o grau de separação entre nós em uma rede será necessário recorrer a alguns trabalhos publicados.

Milgram (1967) apresenta o "problema do mundo pequeno"(small world). O modelo do "mundo pequeno"considera que cada ator em uma rede, pode encontrar outro ator com seis passos em média os "seis graus de separação".

De acordo com Watts e Strogatz (1998) o grau de separação é a distância do menor caminho percorrido para interligar um nó a outro em uma rede. Os experimentos de Watts e Stogatz mostram que em redes nas quais os comprimentos dos caminhos são pequenos e o coeficiente de aglomeração é muito maior do que os das redes randômicas com o mesmo número de nós, são chamadas de redes de mundo pequeno.

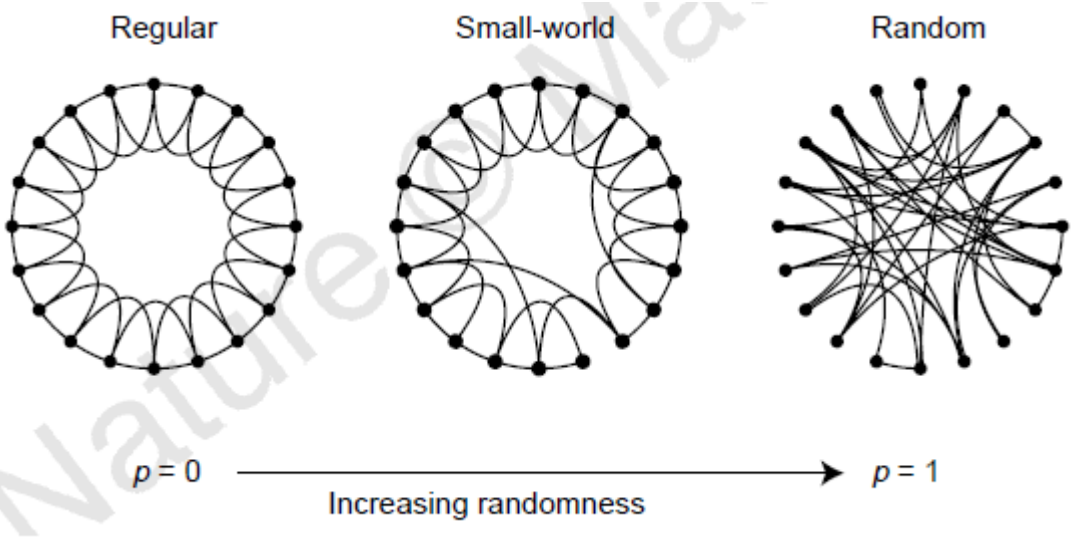

Figura 4 – Representação das redes. A redes de mundo pequeno estão entre as redes regulares de caminhos uniformes e sequanciais e as redes randomicas de caminhos aleatórios

FONTE: (WATTS; STROGATZ, 1998)

Para Balancieri et al. (2005), na linguagem das análises de redes sociais, as pessoas ou os grupos são chamados de "atores", e as conexões, de "ligações". Ambos podem ser definidos em diferentes caminhos, dependendo da questão de interesse. Um ator pode ser uma única pessoa, um grupo ou uma empresa. Newman (2001) apresentam um grafo da rede de colaboração entre Duncan Watts e Albert Barabási, mostrando os caminhos que são possíveis de percorrer para que se conecte os dois atores na rede.

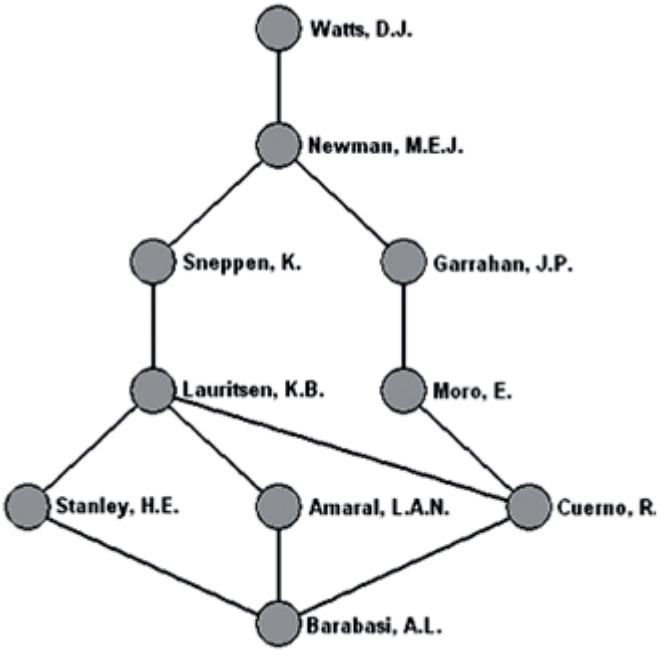

Figura 5 – Rede de colaboração científica entre Dunca Watts Albert Barabási

FONTE: (NEWMAN, 2001)

Redes de mundo pequeno tendem a ser livre de escala ao seguirem a lei de potencia

$$P(x) = ax^k \tag{2.1}$$

. Na natureza existem fatores que se apresentão seguindo a caracterísitca de ser livre de escala. A frenquência com que ocorrem os terremotos pode ser um exemplo. A grande maioria dos terremotos são de pequena intensidade e poucos são de grande intensidade, possuindo os dados da ocorrência de terremotos, pode-se escreve-los em um gráfio que será traduzido pela lei de potência.

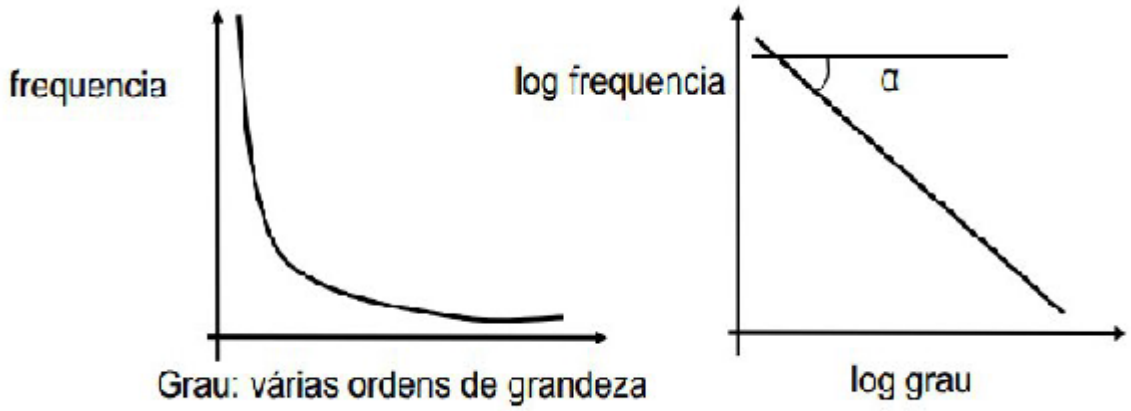

Figura 6 – Graficos da lei de potencia

Ao linearizar utilizando a forma logarítimica da equação, é possível obter a inclinação da reta resultante do gráfico. Kunegis (2011) analisa seis redes sociais e mostra que três delas possuem características de mundo pequeno ao serem consideradas livres de escala. A análise levou em conta a equação da lei de pontência escrita de forma logarítima e mostra em quais redes foi possível obter a inclinação da reta resultante do processo de linearização

$$log(P(x)) = k.log(x) + log(a) \tag{2.2}$$

.

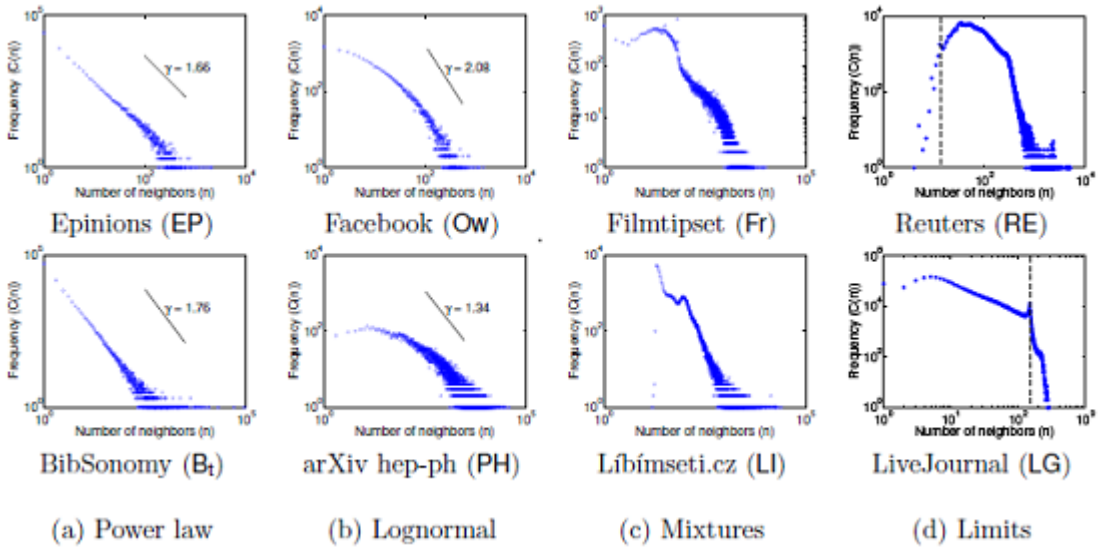

Figura 7 – Análise de redes sociais quanto a caracterísitca de serem livres de escala

FONTE: (KUNEGIS, 2011)

Na predição de links os graus de separação entre os nós é um dificultador. A característica da rede em ser livre de escala ajuda na acurácia da predição de links pelo modelo da evolução espectral de acordo com os esperimentos realizados por Kunegis (2011). Quando o grau de separação é maior do que dois a tarefa de se prever links torna-se mais complexa. Métodos de preedição baseados em vizinhaça comum utilizam a característica de interligação dos nós, o que pode negligenciar as conexões de grau maiores do que dois.

Utilizar a matriz de adjacência da rede analisada pode ser uma ferramenta interessante para se chegar a prever links em uma rede de larga escala, o olhar nessa estratégia estará voltado para a estrutura da rede e não para as características dos nós. Esse é o embasamneto da teoria da evolução espectral (KUNEGIS, 2011).

2.5 A Teoria da Evolução Espectral

No século XVIII um problema proposto por Euler chamou a atenção. A cidade de Konigsberg, um antigo território da Prússia, era dividida por duas ilhas e cortadas por um rio. A interligação entre as duas ilhas era realizada por sete pontes. O problema era encontrar uma maneira de percorrer cada uma das sete pontes de uma só vez de modo a voltar ao ponto de partida. Euler demonstrou que era impossível tal feito, se a quantidade de pontes continuasse a ser um número ímpar. Para provar sua hipótese Euler, institivamente, considerou cada extremidade das pontes os nós e a ponte a interligação, desenhando possivelmente o primeiro grafo da história.

Matematicamente, uma rede pode ser representada por um grafo. A representação matemática de um grafo pode ser dada por:

$$G = (V, E) \tag{2.3}$$

onde V é o conjunto de vértices e E é o conjunto de arestas. Se i e j pertencem a V, i e j são dois vértices e *(i,j)*, representa uma aresta, uma conexão entre os vértices.

Outra forma de representar uma rede é construindo a matriz de adjacência de seu grafo. O método consiste em informar de forma numérica aonde existem conexões entre os nós. Seja G o grafo de uma rede e i, j um dos pares de nós interligados da rede, a representação para essa interligação na matriz e dada pelo número 1, indicando que existe uma conexão.

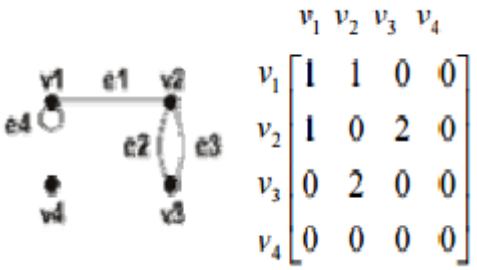

Figura 8 – Matriz de adjacência do grafo contendo 4 nós. A quantidade de conexões são representadas nas intercessões das linhas com as colunas

FONTE: (PEREIRA; CÂMARA, 2011)

O uso da matriz de adjacência de um grafo é a base da teoria espectral dos grafos apresentada por Chung (1997). A teoria espectal dos grafos pode ser utilizada para estudar propriedades da rede como conectividade, centralidade, e o agrupamento. Apoiado nessa idéia Kunegis (2011) apresenta o modelo da evolução espectral e o define como: "uma rede que evolui ao longo do tempo é dito seguindo o modelo da evolução espectral quando o espectro da rede mostra que seus autovetores permanecem aproximadamente contantes". Kunegis (2011) ainda reforça que para o modelo da evolução espectral ser válido o conjuto de autovalores deve ser correspondente aos mesmos autovetores.

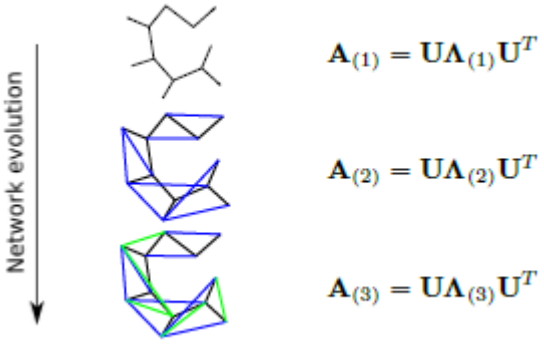

Figura 9 – Representação de uma rede evoluindo seguindo o modelo da evolução espectral, decompondo a matriz de adjacência do grafo em seus autovalores e autovetores

FONTE: (KUNEGIS; FAY; BAUCKHAGE, 2010)

2.6 Autovalores e Autovetores: o formalismo matemático da teoria da evolução espectral

Seja um vetor **x** pertencente aos conjuntos do números reais R, a norma do vetor pode ser definida como:

$$||x|| = \sqrt{\sum_{i=1}^{n} x^2} \tag{2.4}$$

. Se a norma do vetor **x** é igual a 1, pode-se dizer que **x** é um vetor unitário. O produto de dois vetores **x**, **y** pode ser escrito como:

$$x.y = \sum_{i=1}^{n} x.y \tag{2.5}$$

. Se $x.y = 0$, produto escalar nulo, **x** e **y** são ditos vetores ortogonais.

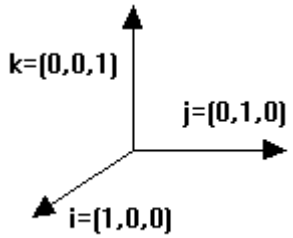

Figura 10 – Vetores ortogonais em três dimensões, o produto escalar entre eles é nulo.

Considerando uma matriz **X** pertencente ao conjuto dos numeros reais R, os componentes x^n representando as linhas e os componentes j^n representando as colunas, escrevem a matriz como X_{ij}. Uma matriz é dita quadrada quando o número de linhas é igual ao númeor de colunas $m = n$. A norma da matriz **X** é definida por:

$$||X|| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2} \tag{2.6}$$

Seja **X** uma matriz m x n não necessariamente simétrica ou quadrática, pode-se escrever a equação dos autovetores da matriz como $(A-\lambda I)v = 0$, atribuindo uma solução não trivial para equação onde $v \neq 0$, a equação pode ser escrita da seguinte forma:

$$Av = \lambda v \tag{2.7}$$

O vetor v é chamado autovetor da equação.

Se a Matriz A é quadrática simétrica, a decomposição em todos os seus autovalores por ser escrita segundo Chung (1997) por:

$$A = U\Lambda U^T \tag{2.8}$$

U é a matriz ortogonal de A e Λ é a matriz diagonal. Combinando as equações 2.6 e 2.7, conclui-se que $UU^T = I$. A matriz ortogonal U multiplicada pela matriz transposta U^T é igual a matriz identidade I.

A representação gráfica da teoria da evolução espectral de Kunegis (2011) pode ser ilustrada pela figura 9. O autovetor é dito constante quando ele cresce dentro do mesmo subespaço, não alterando a sua direção mesmo que os autovalores sejam alterados. O sinal do autovalor pode até mesmo ser negativo. Um autovetor é um vetor em que sua direção não é alterada mesmo que o vetor original passe por uma transformação linear. Ao olhar para equação 2.6 verifica-se que a transformação linear foi causada pelo multiplicador λ. O multiplicador é denominado autovalor porque o subespaço vetorial após a transformação linear não foi alterado.

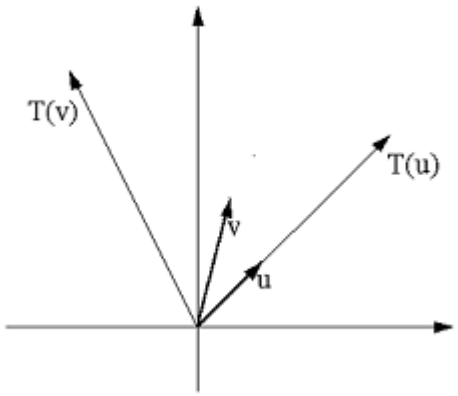

Figura 11 – Representação dos autovetores. u é autorvetor de T, equanto v não. Logo u e v representam diferentes subespaços

3 Metodologia

De acordo com Rodrigues (2006) a metodologia científica define os tipos de pesquisa classificadas quanto: a área da ciência, a natureza, aos objetivos, ao objeto, aos procedimentos e a forma de abordagem. As seções desse capítulo apresentará cada uma das características presentes no projeto de pesquisa.

3.1 A Área da Ciência

A pesquisa quanto a área da ciência será considerada *Pesquisa Prática* (RODRIGUES, 2006). De acordo com Candy () a pesquisa prática é direcionada a fim de se obter novos conhecimentos através da prática e dos resultados dessa prática. Alegações de originalidade e contribuições para o conhecimento podem ser demonstrados através de resultados que podem ser artefactos como imagens, músicas, desenhos, modelos, meios digitais ou outros resultados, tais como análise de performances, enquanto o significado e o contexto das afirmações são descritos em palavras, uma compreensão completa só pode ser obtida com referência direta aos resultados.

3.2 A Natureza

A natureza do trabalho é tratada como *Resumo de Assunto*. Para Severino et al. (2000) o resumo de assunto, consiste de uma pesquisa fundamentada em trabalhos mais avançados publicados por autoridades no assunto. A originalidade do trabalho pode ser dispensada, mas não o rigor científico, não se limita a ser uma simples cópia de idéias, as principais qualidades de um trabalho classificado como *Resumo de Assunto* são: análise e interpretação de fatos e idéias, metodologia adequada, originalidade do ponto de vista do enfoque do tema.

3.3 Os Objetivos

Os objetivos da pesquisa conduzem-na para ao tipo de *Pesquisa Exploratória*. "Estas pesquisas têm como objetivo proporcionar maior familiaridade com o problema, com vistas a tomá-lo mais explícito ou a constituir hipóteses. Pode-se dizer que estas pesquisas têm como objetivo principal o aprimoramento de idéias ou a descoberta de intuições. Seu planejamento é portanto bastante flexível, de modo que possibilite a consideração dos mais variados aspectos relativos ao fato estudado."(GIL, 2002).

3.4 Os Procedimentos

Quanto aos procedimentos a pesquisa será realizada por *Fontes*. Nessas fontes estão as pesquisas bibliográficas. As pesquisas bibliográficas serão realizada em bases secundárias: artigos publicados, livros, e teses.

3.5 O Objeto

Quanto ao objeto definiu-se por realizar a Pesquisa Bibliográfica e a Prova de Conceito .

"A pesquisa bibliográfica é desenvolvida com base em material já elaborado, constituído principalmente de livros e artigos científicos. Embora em quase todos os estudos seja exigido algum tipo de trabalho dessa natureza, há pesquisas desenvolvidas exclusivamente a partir de fontes bibliográficas. Boa parte dos estudos exploratórios pode ser definida como pesquisas bibliográficas."(GIL, 2002). Ainda segundo Gil (2002) a principal vantagem da pesquisa bibliográfica reside no fato de permitir ao investigador a cobertura de uma gama de fenômenos muito mais ampla do que aquela que poderia pesquisar diretamente. Essa vantagem torna-se particularmente importante quando o problema de pesquisa requer dados muito dispersos pelo espaço. Por exemplo, seria impossível a um pesquisador percorrer todo o território brasileiro em busca de dados sobre população ou renda per capita; todavia, se tem a sua disposição uma bibliografia adequada, não tem maiores obstáculos para contar com as informações requeridas. A pesquisa bibliográfica também é indispensável nos estudos históricos. Em muitas situações, não há outra maneira de conhecer os fatos passados se não com base em dados bibliográficos.

A Prova de Conceito, é um termo utilizado para denominar um modelo prático que possa provar um conceito teórico. A prova de conceito pode utilizar um protótipo de um software que foi desenvolvido baseado em um conceito teórico. A performance do software será o resultado da prova de conceito.

3.6 Etapas da Pesquisa

Importante ressaltar que o projeto de dissertação está atrelado ao projeto entitulado *Uso da Análise Espectral para Predição de Links em uma Rede de Coautoria* apresentado ao Edital ProPIC 01/2012 da Universidade FUMEC. O projeto foi elaborado pelo professor orientador desse projeto de dissertação, Orlando Abreu Gomes, e o autor desse projeto de dissertação é participante do grupo de pesquisa como bolsista estudante de mestrado. O projeto tem o obejtivo de avaliar a predição de links na rede de coautoria dos artigos publicados nos encontros anuais da AMPAD (Associação Nacional de Pós-Graduação e Pesquisa em Administração).

As etapas propostas para se atingir os objetivos são:

1 - A pesquisa bibliográfica irá buscar trabalhos relacionados a utilização de técnicas de predição de links em redes sociais e de coautoria. Pretende-se com essa pesquisa criar um mapeamento da evolução das teorias de acordo com a data de publicação dos trabalhos, esse mapeamento proporcionará classificar os trabalhos. No capítulo 2 desse projeto de dissertação já estão apresentados alguns resultados. Sem a revisão de literatura pronta, a dificuldade para escrita do projeto seria aumentada.

2 - A base da "Plataforma Lattes"é extensa e para um trabalho específico como realizar a predição de links em sua rede coautoria, faz-se necessário uma mineração dos dados. A mineração dos dados terá o objetivo de filtrar a base e torna-la passível de aplicação de um software de mineração. Na filtragem da base, considera-se retirar nomes de autores de forma duplicada, caracteres não identificados e dados irrelevantes para se construir o espectro da rede.

3 - Contrução dos grafos da rede, representando sua evolução ao longo do tempo. Os grafos seram construídos tendo como parâmetro de particionamento da rede os anos de evolução. Se a base minerada está composta pelos trabalhos publicados entre os anos de 2000 a 2013, o objetivo será construir 14 grafos, um para cada ano.

4 - Construção das matrizes de adjacência dos respectivos grafos. Será desenvolvido um software para tal finalidade.

5 - Aplicação da teoria da evolução espectral nas matrizes de adjacência, levando em conta os autovalores e os autovetores. Será utilizado um sofware específico para cálculos matemáticos, o *"Octave"* (https://www.gnu.org/software/octave/).

6 - Obtenção do espectro da rede e plotagem em um gráfico. A plotegem no gráfico permitirá apontar se a evolução da rede está ocorrendo dentro do mesmo subespaço vetorial, seguindo a equação 2.6 e a figura 9.

7 - Avaliação dos resultados da aplicação da técnica utilizando os indicadores propostos na teoria de Kunegis (2011).

8 - Extrapolação do espectro da rede para sugerir onde novos links irão surgir.

4 Cronograma

Visando atingir os objetivos propostos apresenta-se um cronograma das atividades. Estas atividades estão ilustradas na tabela 1.

Tabela 1 – Cronograma das atividades

Fase	set/13	out/13	nov/13	dez/13	jan/14	fev/14	mar/14	abr/14
Pesquisa Bibliográfica	x	x	x	x	x	x	x	
Defesa do Projeto								x
Fase	mai/14	jun/14	jul/14	ago/14	set/14	out/14	nov/14	dez/14
Mineração dos Dados	x	x	x					
Construção dos Grafos da rede				x	x			
Aplicação da Teorial				x	x			
Obtenção do Espectro				x	x			
Avaliação dos Resultados						x	x	
Extrapolação do Espectro						x	x	
Redação da Dissertação								x
Defesa da Dissertação								x

Referências

AHMAD, M. A. et al. Link prediction across multiple social networks. In: IEEE. *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*. [S.l.], 2010. p. 911–918. Citado na página 13.

ALLALI, O.; MAGNIEN, C.; LATAPY, M. Link prediction in bipartite graphs using internal links and weighted projection. In: IEEE. *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*. [S.l.], 2011. p. 936–941. Citado na página 13.

BALANCIERI, R. et al. A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na plataforma lattes. *Ciência da Informação*, SciELO Brasil, v. 34, n. 1, p. 64–77, 2005. Citado na página 16.

BRANDÃO, W. C.; PARREIRAS, F. S.; SILVA, A. B. d. O. Redes em ciência da informação: evidências comportamentais dos pesquisadores e tendências evolutivas das redes de coautoria. *Informação & Informação*, Universidade Estadual de Londrina, 2007. Citado na página 14.

CALLIOLI, C.; DOMINGUES, H.; COSTA, R. *Álgebra linear e aplicações*. Atual, 2007. ISBN 9788570562975. Disponível em: <http://books.google.com.br/books?id=AjsRRwAACAAJ>. Citado na página 7.

CANDY, L. Practice based research: A guide. Citado na página 23.

CHANG, C.; YAO, X. Social network link predict based on af model. In: IEEE. *Computer Science and Network Technology (ICCSNT), 2011 International Conference on*. [S.l.], 2011. v. 1, p. 415–418. Citado na página 12.

CHUNG, F. R. *Spectral graph theory*. [S.l.]: American Mathematical Soc., 1997. Citado 2 vezes nas páginas 20 e 21.

DONG, Y. et al. Link prediction based on local information. In: IEEE. *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*. [S.l.], 2011. p. 382–386. Citado na página 12.

FONSECA, A. A. M.; ONEILL, M. M. A revolução tecnológica e informacional eo renascimento das redes. *Revista de Geociências, Niterói, RJ*, v. 2, p. 26–35, 2001. Citado na página 13.

GIL, A. C. Como classificar as pesquisas. _____. *Como elaborar projetos de pesquisa*, v. 4, p. 41–56, 2002. Citado 2 vezes nas páginas 23 e 24.

KASHIMA, H.; ABE, N. A parameterized probabilistic model of network evolution for supervised link prediction. In: *Sixth International Conference on Data Mining, 2006. ICDM '06*. [S.l.: s.n.], 2006. p. 340–349. Citado na página 12.

KITCHENHAM, B. *Procedures for Performing Systematic Reviews*. [S.l.], 2004. Citado na página 10.

KUNEGIS, J. *On the Spectral Evolution of Large Networks*. Tese (Doutorado) — UniversitÃt Koblenz-Landau, Campus Koblenz, 2011. Citado 8 vezes nas páginas 8, 9, 13, 18, 19, 20, 22 e 25.

KUNEGIS, J.; FAY, D.; BAUCKHAGE, C. Network growth and the spectral evolution model. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2010. (CIKM '10), p. 739–748. ISBN 978-1-4503-0099-5. Disponível em: <http://doi.acm.org/10.1145/1871437.1871533>. Citado 5 vezes nas páginas 7, 9, 13, 14 e 20.

KUNEGIS, J.; LOMMATZSCH, A. Learning spectral graph transformations for link prediction. In: ACM. *Proceedings of the 26th Annual International Conference on Machine Learning*. [S.l.], 2009. p. 561–568. Citado na página 13.

Lü, L.; ZHOU, T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, v. 390, p. 1150–1170, 2011. Citado na página 12.

LIBEN-NOWELL, D.; KLEINBERG, J. The link prediction problem for social networks. In: *Proceedings of the Twelfth International Conference on Information and Knowledge Management*. [S.l.]: ACM, 2003. p. 556–559. Citado 2 vezes nas páginas 8 e 12.

MAKREHCHI, M. Social link recommendation by learning hidden topics. In: ACM. *Proceedings of the fifth ACM conference on Recommender systems*. [S.l.], 2011. p. 189–196. Citado na página 12.

MILGRAM, S. The small world problem. *Psychology today*, New York, v. 2, n. 1, p. 60–67, 1967. Citado na página 15.

NEWMAN, M. E. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 98, n. 2, p. 404–409, 2001. Citado 3 vezes nas páginas 14, 16 e 17.

PAPADIMITRIOU, A.; SYMEONIDIS, P.; MANOLOPOULOS, Y. Friendlink: Link prediction in social networks via bounded local path traversal. In: IEEE. *Computational Aspects of Social Networks (CASoN), 2011 International Conference on*. [S.l.], 2011. p. 66–71. Citado na página 15.

PARREIRAS, F. S. et al. Redeci: colaboração e produção científica em ciência da informação no brasil. *Perspectivas em ciência da informação*, SciELO Brasil, v. 11, n. 3, p. 302–317, 2006. Citado na página 14.

PEREIRA, G. M. R.; CÂMARA, M. A. da. *Algumas Aplicações da Teoria dos Grafos*. [S.l.]: Famat em Revista, 2011. Citado na página 20.

RODRIGUES, A. d. J. Metodologia científica. *São Paulo: Avercamp*, v. 222, 2006. Citado na página 23.

RODRIGUEZ, M. G.; ROGATI, M. Bridging offline and online social graph dynamics. In: ACM. *Proceedings of the 21st ACM international conference on Information and knowledge management*. [S.l.], 2012. p. 2447–2450. Citado na página 12.

SANTANA, M. R. C. Redes técnicas: os avatares geográficos da cidade mediada eletronicamente. *Reflexões e Contribuições Geográficas Contemporâneas. Salvador*, 2004. Citado na página 13.

SEVERINO, A. J. et al. *Metodologia do trabalho científico*. [S.l.]: Cortez São Paulo, 2000. Citado na página 23.

SILVA, A. K. A. D.; BARBOSA, R. R.; DUARTE, E. N. Rede social de coautoria em ciência da informação: estudo sobre a área temática de"organização e representação do conhecimento". *Informação & Sociedade: Estudos*, v. 22, n. 2, 2012. Citado na página 14.

SUN, Y. et al. Co-author relationship prediction in heterogeneous bibliographic networks. In: IEEE. *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*. [S.l.], 2011. p. 121–128. Citado na página 15.

TYLENDA, T.; ANGELOVA, R.; BEDATHUR, S. Towards time-aware link prediction in evolving social networks. In: ACM. *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*. [S.l.], 2009. p. 9. Citado na página 12.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. *nature*, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citado 2 vezes nas páginas 15 e 16.